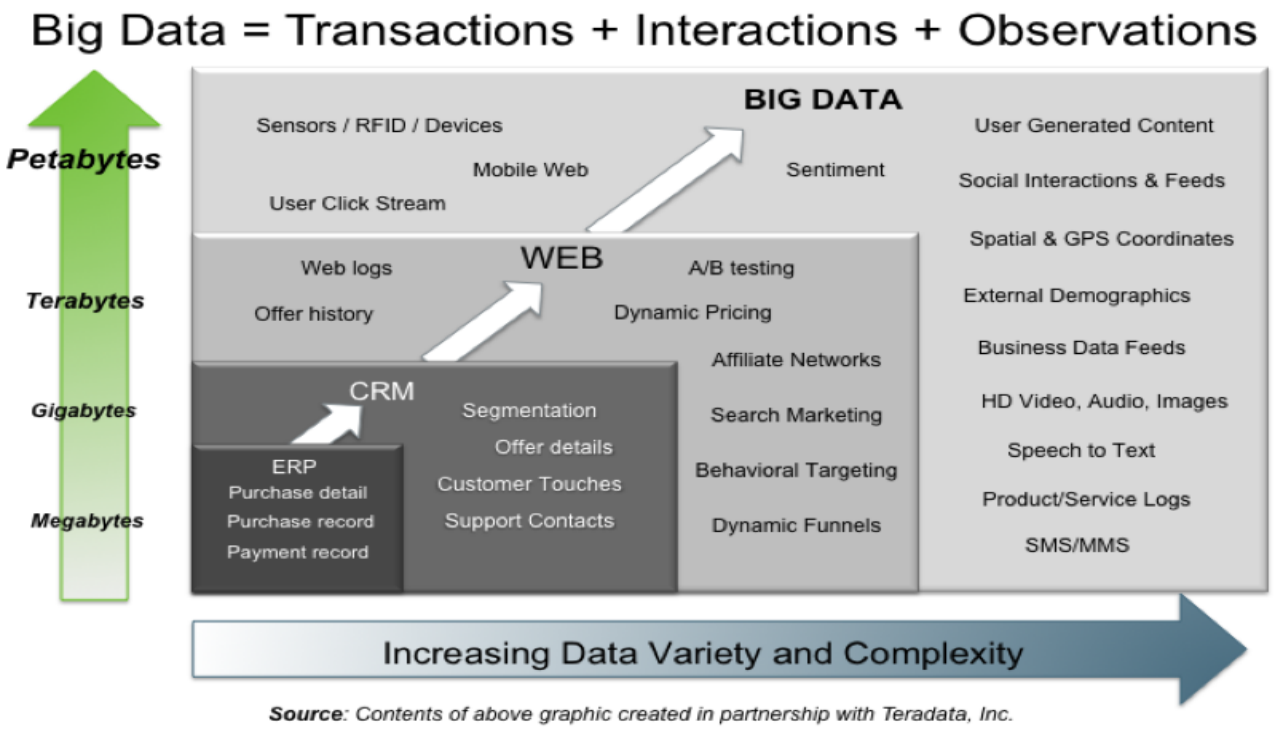
Big Data Analytics

UNIT – I

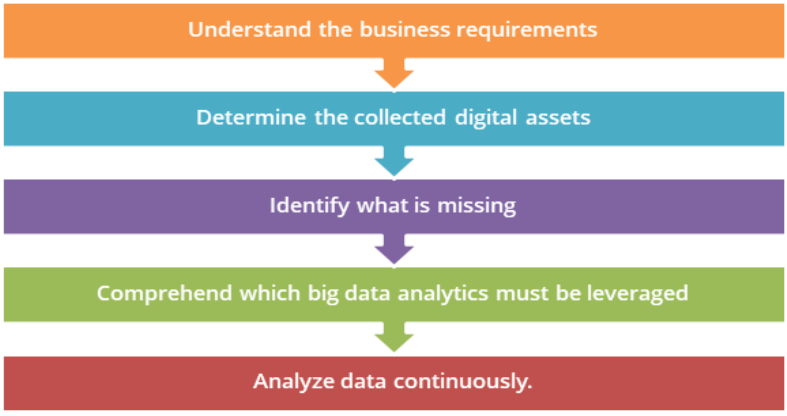
1. Explain The Evolution of BigData.



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

2. Describe the Best Practices for Big data Analytics.

Big data- the word says it all- is an enormous amount of data that gets collected and generated across organizations, social media, Internet, and various other sources. Big data analytics analyses the collected data and find patterns from it. The velocity, veracity, variety, and volume of data lying with organizations must be put to work to gain actionable insights out of the same. Organizations leveraging big data analytics must thoroughly understand the best practices for big data first to be able to use the most relevant data for analysis.



ALL ABOUT BIG DATA: Big data offers countless benefits to several industries, including healthcare, retail, finance, manufacturing, insurance, pension, and many more. But where does all this data come from? Organizations collect and generate a significant amount of data from multiple internal and external sources and it is crucial to manage this data efficiently and securely. This extensive data pouring into an organization is termed as big data. Handling such massive volumes of data with traditional methods is tedious; hence big data analysis came into existence. It is imperative to analyse digital assets in the organization thoroughly to get an insight into the effectiveness of any existing processes and practices. Big data analytics help find patterns in the collected data sets, which allows business users to identify and analyze emerging market trends. Moreover, big data analytics helps various industries to find new opportunities and improve in areas where they lack.

BEST PRACTICES FOR BIG DATA

Now, with the knowledge of what is big data and what it offers, organizations must know how analytics must be practiced to make the most of their data. The list below shows five of the best practices for big data:

1. UNDERSTAND THE BUSINESS REQUIREMENTS

Analyzing and understanding the business requirements and organizational goals is the first and the foremost step that must be carried out even before leveraging big data analytics into your projects. The business users must understand which projects in their company must use big data analytics to make maximum profit.

2. DETERMINE THE COLLECTED DIGITAL ASSETS

The second best big data practice is to identify the type of data pouring into the organization, as well as, the data generated in-house. Usually, the data collected is disorganized and in varying formats. Moreover, some data is never even exploited (read dark data), and it is essential that organizations identify this data too.

3. IDENTIFY WHAT IS MISSING

The third practice is analyzing and understanding what is missing. Once you have collected the data needed for a project, identify the additional information that might be required for that particular project and where can it come from. For instance, if you want to leverage big data analytics in your organization to understand your employee's well-being, then along with information such as login logout time, medical reports, and email reports, we need to have some additional information about the employee’s, let’s say, stress levels. This information can be provided by co-workers or leaders.

4. **COMPREHEND WHICH BIG DATA ANALYTICS MUST BE LEVERAGED**

After analyzing and collecting data from different sources, it's time for the organization to understand which big data technologies, such as predictive analytics, stream analytics, data preparation, fraud detection, sentiment analysis, and so on can be best used for the current business requirements. For instance, big data analytics helps the HR team in companies for the

recruitment process to identify the right talent faster by collaborating the social media and job portals using predictive and sentiment analysis.

5. ANALYZE DATA CONTINUOUSLY

This is the final best practice that an organization must follow when it comes to big data. You must always be aware of what data is lying with your organization and what is being done with it. Check the health of your data periodically to never miss out on any important but hidden signals in the data. Before implementing any new technology in your organization, it is vital to have a strategy to help you get the most out of it. With adequate and accurate data at their disposal, companies must also follow the above mentioned big data practices to extract value from this data.

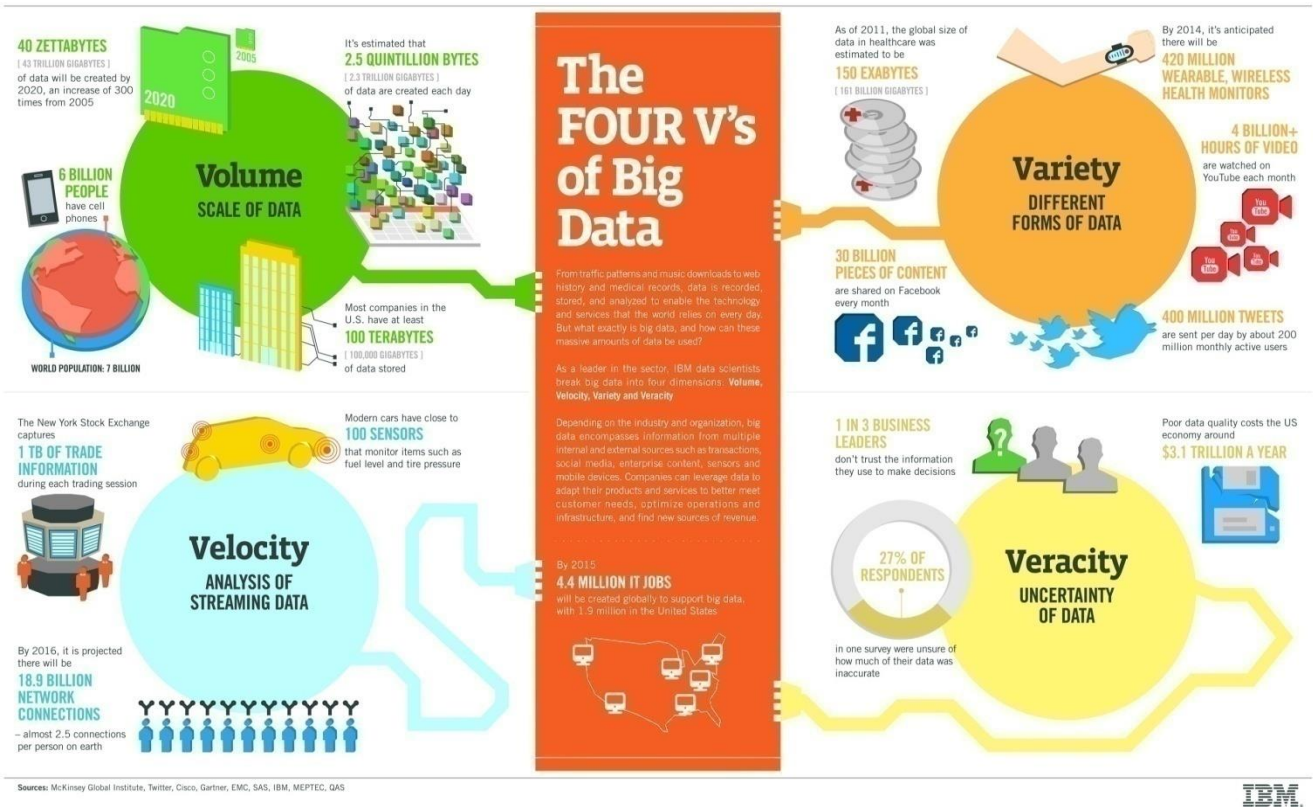
3. Define Elements OF Big Data or Characteristics Of Big Data.

(i) Volume – The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, **'Volume'** is one characteristic which needs to be considered while dealing with Big Data.

(ii) Variety – Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analysing data.

(iii) Velocity – The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.
Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

(iv) Veracity / Variability – This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.



4 Issues and Challenges of Big Data.

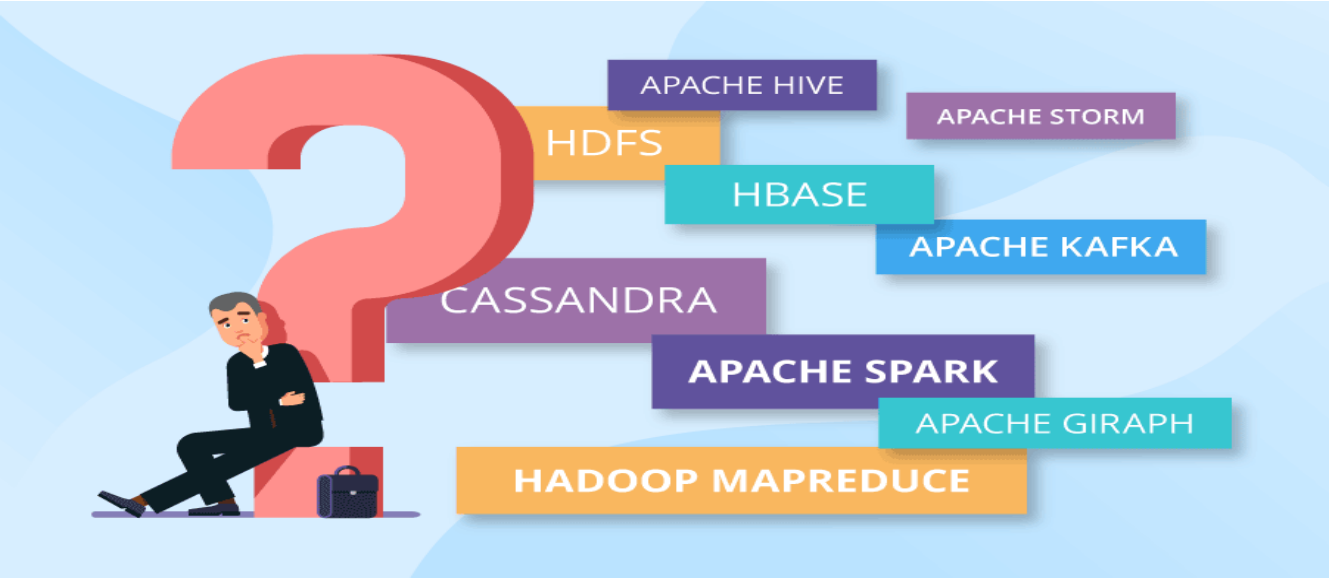
Challenge #1: Insufficient understanding and acceptance of big data

Oftentimes, companies fail to know even the basics: what big data actually is, what its benefits are, what infrastructure is needed, etc. Without a clear understanding, a big data adoption project risks to be doomed to failure. Companies may waste lots of time and resources on things they don't even know how to use. And if employees don't understand big data's value and/or don't want to change the existing processes for the sake of its adoption, they can resist it and impede the company's progress.

Solution:

Big data, being a huge change for a company, should be accepted by top management first and then down the ladder. To ensure big data understanding and acceptance at all levels, IT departments need to organize numerous trainings and workshops. To see to big data acceptance even more, the implementation and use of the new big data solution need to be monitored and controlled. However, top management should not overdo with control because it may have an adverse effect.

Challenge #2: Confusing variety of big data technologies

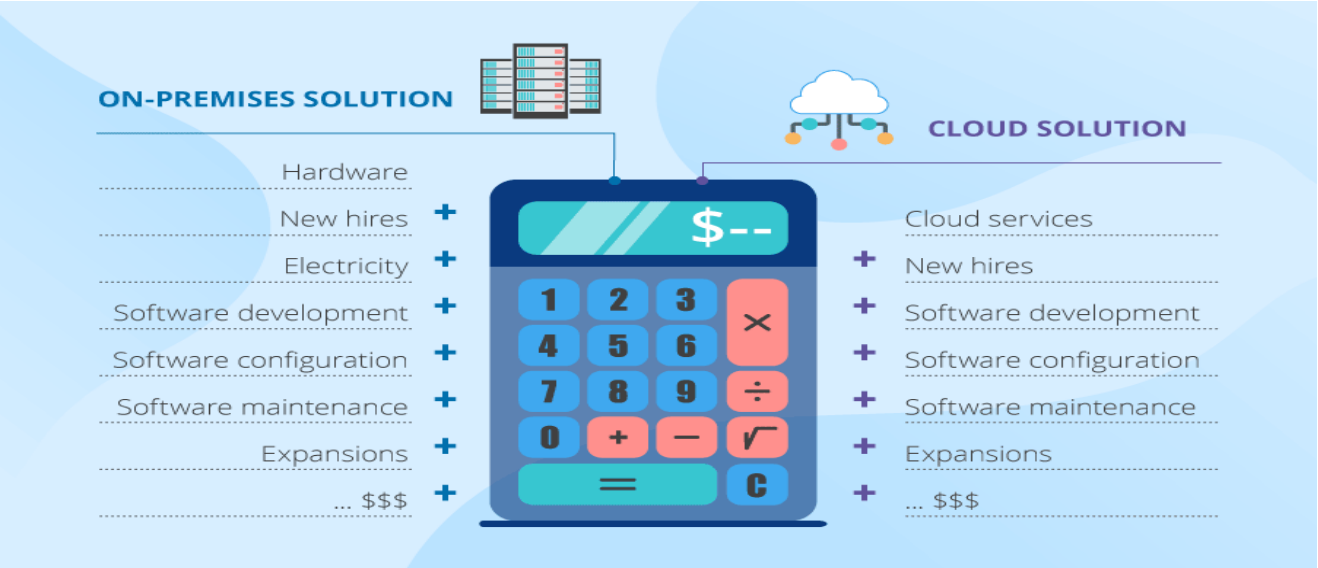


It can be easy to get lost in the variety of big data technologies now available on the market. Do you need Spark or would the speeds of Hadoop MapReduce be enough? Is it better to store data in Cassandra or HBase? Finding the answers can be tricky. And it's even easier to choose poorly, if you are exploring the ocean of technological opportunities without a clear view of what you need.

Solution:

If you are new to the world of big data, trying to seek professional help would be the right way to go. You could hire an expert or turn to a vendor for big data consulting. In both cases, with joint efforts, you'll be able to work out a strategy and, based on that, choose the needed technology stack.

Challenge #3: Paying loads of money



Big data adoption projects entail lots of expenses. If you opt for an on-premises solution, you'll have to mind the costs of new **hardware**, new **hires** (administrators and developers), **electricity** and so on. Plus: although the needed frameworks are open-source, you'll still need to pay for the development, setup, configuration and maintenance of

new **software**. If you decide on a cloud-based big data solution, you'll still need to **hire staff** (as above) and pay for **cloud services**, big data solution **development** as well as setup and maintenance of needed **frameworks**. Moreover, in both cases, you'll need to allow for future **expansions** to avoid big data growth getting out of hand and costing you a fortune.

Solution:

The particular salvation of your company's wallet will depend on your company's specific technological needs and business goals. For instance, companies who want flexibility benefit from cloud. While companies with extremely harsh security requirements go on-premises. There are also hybrid solutions when parts of data are stored and processed in cloud and parts – on-premises, which can also be cost-effective. And resorting to data lakes or algorithm optimizations (if done properly) can also save money:

1. **Data lakes** can provide cheap storage opportunities for the data you don't need to analyze at the moment.
2. **Optimized algorithms**, in their turn, can reduce computing power consumption by 5 to 100 times. Or even more.

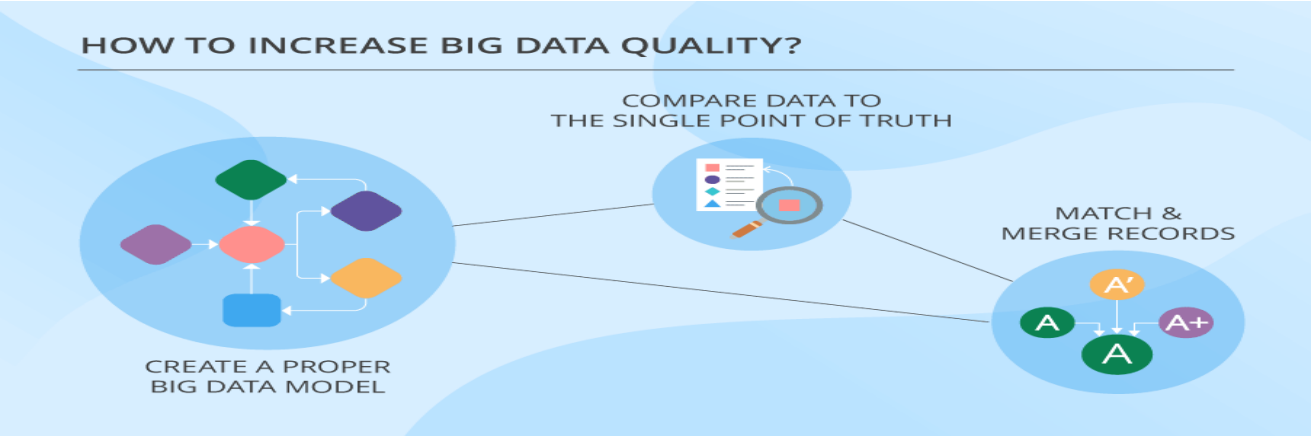
All in all, the key to solving this challenge is properly analyzing your needs and choosing a corresponding course of action.

Challenge #4: Complexity of managing data quality

Data from diverse sources: Sooner or later, you'll run into the problem of data integration, since the data you need to analyze comes from diverse sources in a variety of different formats. For instance, ecommerce companies need to analyze data from website logs, call-centers, competitors' website 'scans' and social media. Data formats will obviously differ, and matching them can be problematic. For example, your solution has to know that skis named SALOMON QST 92 17/18, Salomon QST 92 2017-18 and Salomon QST 92 Skis 2018 are the same thing, while companies ScienceSoft and Sciencesoft are not.

Unreliable data: Nobody is hiding the fact that big data isn't 100% accurate. And all in all, it's not that critical. But it doesn't mean that you shouldn't at all control how reliable your data is. Not only can it contain wrong information, but also duplicate itself, as well as contain contradictions. And it's unlikely that data of extremely inferior quality can bring any useful insights or shiny opportunities to your precision-demanding business tasks.

Solution:



There is a whole bunch of techniques dedicated to cleansing data. But first things first. Your big data needs to have a proper model. Only after creating that, you can go ahead and do other things, like:

- Compare data to the single point of truth (for instance, compare variants of addresses to their spellings in the postal system database).
- Match records and merge them, if they relate to the same entity.

But mind that big data is never 100% accurate. You have to know it and deal with it, which is something this article on big data quality can help you with.

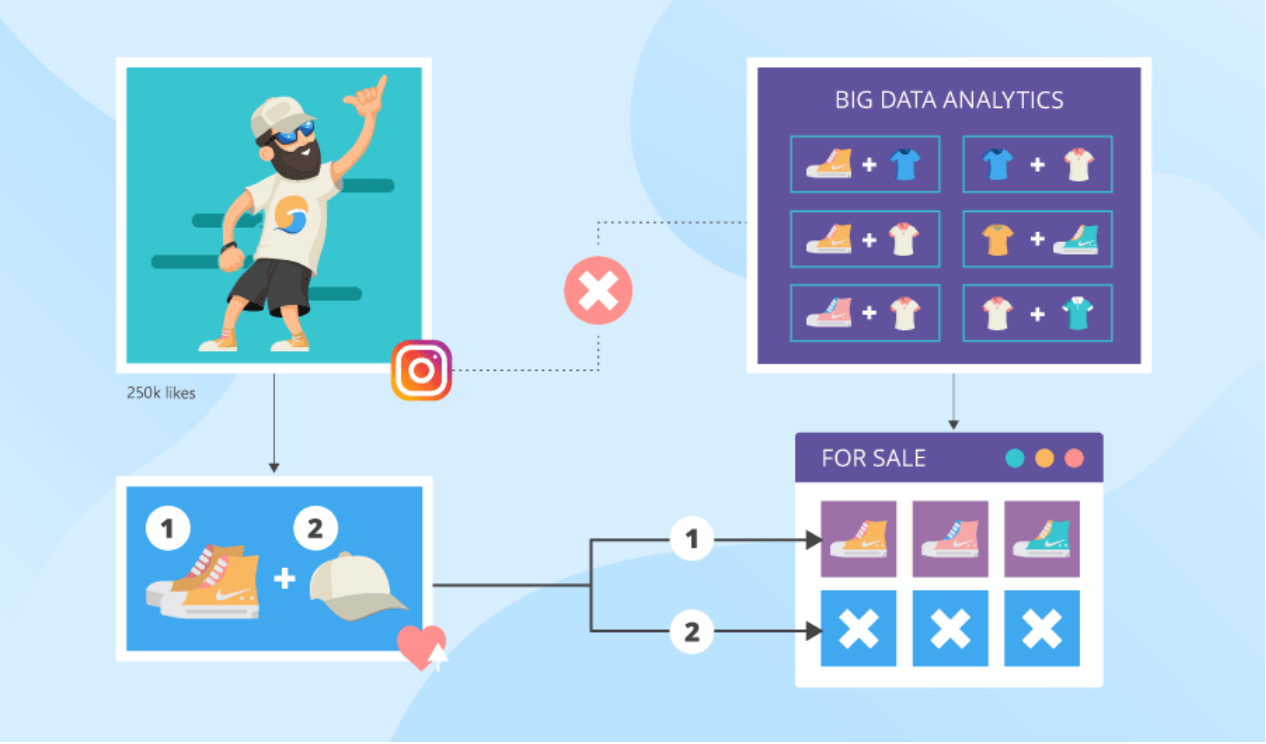
Challenge #5: Dangerous big data security holes

Security challenges of big data are quite a vast issue that deserves a whole other article dedicated to the topic. But let's look at the problem on a larger scale. Quite often, big data adoption projects put security off till later stages. And, frankly speaking, this is not too much of a smart move. Big data technologies do evolve, but their security features are still neglected, since it's hoped that security will be granted on the application level. And what do we get? Both times (with technology advancement and project implementation) big data security just gets cast aside.

Solution:

The precaution against your possible big data security challenges is putting security first. It is particularly important at the stage of designing your solution's architecture. Because if you don't get along with big data security from the very start, it'll bite you when you least expect it.

Challenge #6: Tricky process of converting big data into valuable insights



Here's an example: your super-cool big data analytics looks at what item pairs people buy (say, a needle and thread) solely based on your historical data about customer behavior. Meanwhile, on Instagram, a certain soccer player posts his new look, and the two characteristic things he's

wearing are white Nike sneakers and a beige cap. He looks good in them, and people who see that want to look this way too. Thus, they rush to buy a similar pair of sneakers and a similar cap. But in your store, you have only the sneakers. As a result, you lose revenue and maybe some loyal customers.

Solution:

The reason that you failed to have the needed items in stock is that your big data tool doesn't analyze data from social networks or competitor's web stores. While your rival's big data among other things does note trends in social media in near-real time. And their shop has both items and even offers a 15% discount if you buy both. The idea here is that you need to create a proper system of factors and data sources, whose analysis will bring the needed insights, and ensure that nothing falls out of scope. Such a system should often include external sources, even if it may be difficult to obtain and analyse external data.

Challenge #7: Troubles of up scaling

The most typical feature of big data is its dramatic ability to grow. And one of the most serious challenges of big data is associated exactly with this. Your solution's design may be thought through and adjusted to up scaling with no extra efforts. But the real problem isn't the actual process of introducing new processing and storing capacities. It lies in the complexity of scaling up so, that your system's performance doesn't decline and you stay within budget.

Solution:

The first and foremost precaution for challenges like this is a decent architecture of your big data solution. As long as your big data solution can boast such a thing, less problems are likely to occur later. Another highly important thing to do is designing your big data algorithms while keeping future up scaling in mind. But besides that, you also need to plan for your system's maintenance and support so that any changes related to data growth are properly attended to. And on top of that, holding systematic performance audits can help identify weak spots and timely address them.

5 Stages of Big Data Analytical Evolution.

The process of dealing with big data is quite different from handling traditional data. Big Data processing consists of collecting storing, organizing, analyzing and extracting hidden information for decision making.

1. Data Collection: This is the first stage which involves the collection of web data, log data, structured and unstructured data from several types of data sources, like mobile devices, sensor devices, social media.

2. Storing: In this stage the collected data has to be stored into distributed database systems and servers. Introduction to NOSQL facilitated to store big data. Since NOSQL does not have any fixed schema and there is no relationship between entities it is used to store dynamic and un structured data.

3. Data Organization and Analysis: In this stage, data is arranged and organized as structured, unstructured and semi-unstructured data, in order to access and analyzed. After the data is arranged and organized, the analysis stage is applied. Analyzing large data set involves more complexities and computations. More research and survey is going on to find the algorithm and mathematical model to minimize the computational and storage cost. The extracted hidden information will be useful for the Industries, Academicians and the Government to make necessary action and decision . The infrastructure needed for big data should be highly scalable, support statistical analytics and data mining and based on analytical model automated decision should be made in quick time.

4. Data Visualization: Once the information has been carried out from data, it has to be represented in a visualized manner. The representation is generally done using Data Visualization tools that enable decision makers to grasp difficult concepts and pattern easily.

6 State of the Practice in Analytics.

Current business problems provide many opportunities for organizations to become more analytical and data driven

Business Drivers for Advanced Analytics

Business Driver	Examples
Optimize business operations	Sales, pricing, profitability, efficiency
Identify business risk	Customer churn, fraud, default
Predict new business opportunities	Upsell, cross-sell, best new customer prospects
Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Basel II-III, Sarbanes-Oxley (SOX)

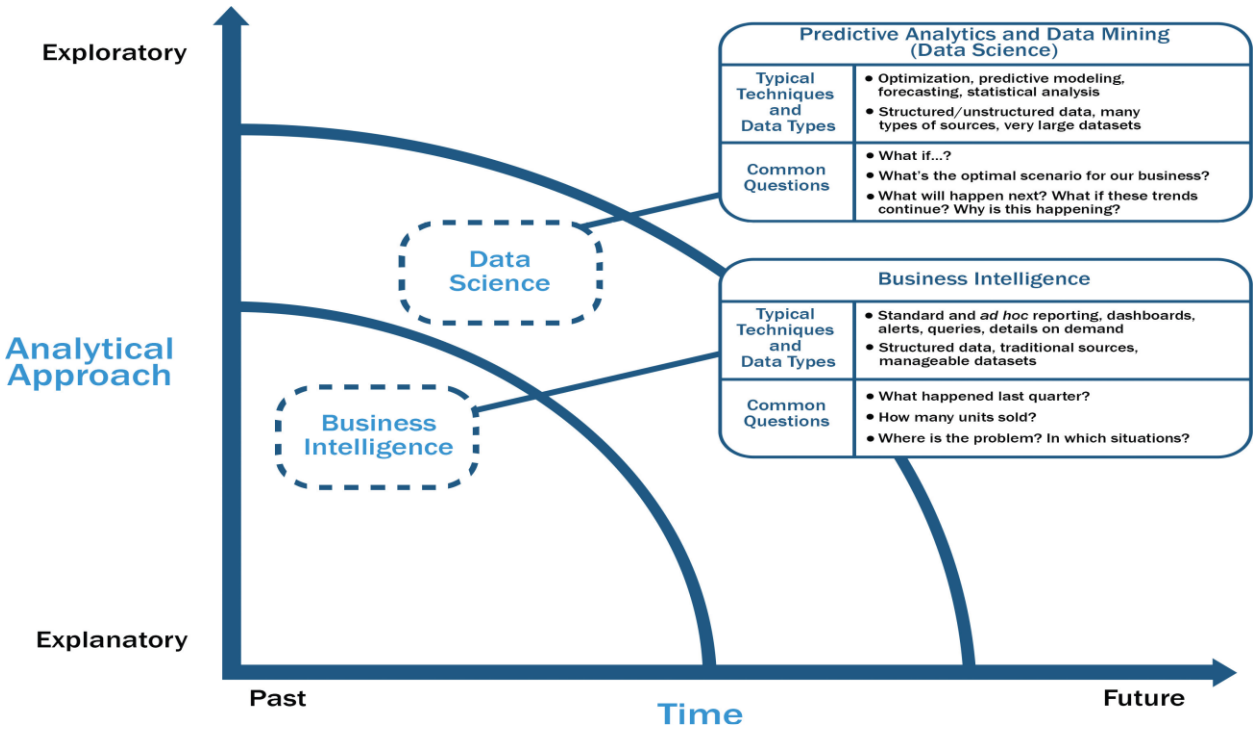
The four categories of common business problems that organizations contend with where they have an opportunity to leverage advanced analytics to create competitive advantage. Rather than only performing standard reporting on these areas, organizations can apply advanced analytical techniques to optimize processes and derive more value from these common tasks. The first three examples do not represent new problems. Organizations have been trying to reduce customer churn, increase sales, and cross-sell customers for many years. What is new is the opportunity to fuse advanced analytical techniques with Big Data to produce more impactful analyses for these traditional problems. The last example portrays emerging regulatory requirements. Many compliance and regulatory laws have been in existence for decades, but additional requirements are added every year, which represent additional complexity and data requirements for organizations. Laws related to anti-money laundering

(AML) and fraud prevention require advanced analytical techniques to comply with and manage properly.

BI Versus Data Science:

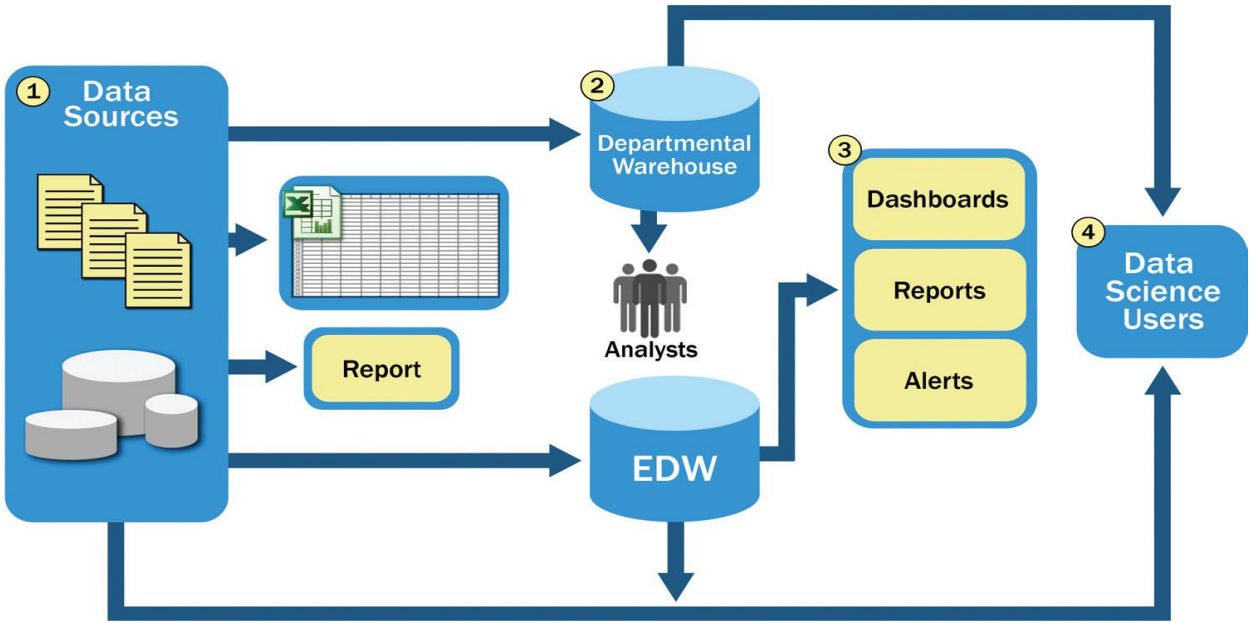
BI and Data Science are group of analytical techniques

- BI tends to provide reports, dashboards, and queries on business questions for the current period or in the past.
- Data Science tends to use disaggregated data in a more forward-looking, exploratory way, focusing on analysing the present and enabling informed decisions about the future.
- BI systems make it easy to answer questions related to quarter-to-date revenue, progress toward quarterly targets, and understand how much of a given product was sold in a prior quarter or year.
- Rather than aggregating historical data to look at how many of a given product sold in the previous quarter, a team may employ Data Science techniques such as time series analysis, Time Series Analysis," to forecast future product sales and revenue more accurately than extending a simple trend line
- These questions tend to be closed-ended and explain current or past behavior, typically by aggregating historical data and grouping it in some way.
- Data Science tends to be more exploratory in nature and may use scenario optimization to deal with more open-ended questions.
- BI provides hindsight and some insight and generally answers questions related to "when" and "where" events occurred.
- Data Science provides insight into current activity and foresight into future events, while generally focusing on questions related to "how" and "why" events occur.
- BI problems tend to require highly structured data organized in rows and columns for accurate Reporting
- Data Science projects tend to use many types of data sources, including large or unconventional datasets.
- Depending on an organization's goals, it may choose to embark on a BI project if it is doing reporting, creating dashboards, or performing simple visualizations,
- It may choose Data Science projects if it needs to do a more sophisticated analysis with disaggregated or varied datasets.



Current Analytical Architecture:

- Data Science projects need workspaces that are purpose-built for experimenting with data, with flexible and agile data architectures.
- Most organizations still have data warehouses that provide excellent support for traditional reporting and simple data analysis activities but unfortunately have a more difficult time supporting more robust analyses.
- Figure shows a typical data architecture and several of the challenges it presents to data scientists and others trying to do advanced analytics.



1. For data sources to be loaded into the data warehouse, data needs to be well understood, structured, and normalized with the appropriate data type definitions. Although this kind of centralization enables security, backup, and failover of highly critical data, it also means that data typically must go through significant preprocessing and checkpoints before it can enter this sort of controlled environment, which does not lend itself to data exploration and iterative analytics.

2. As a result of this level of control on the EDW, additional local systems may emerge in the form of departmental warehouses and local data marts that business users create to accommodate their need for flexible analysis. These local data marts may not have the same constraints for security and structure as the main EDW and allow users to do some level of more in-depth analysis. However, these one-off systems reside in isolation, often are not synchronized or integrated with other data stores, and may not be backed up.

3. Once in the data warehouse, data is read by additional applications across the enterprise for BI and reporting purposes. These are high-priority operational processes getting critical data feeds from the data warehouses and repositories.

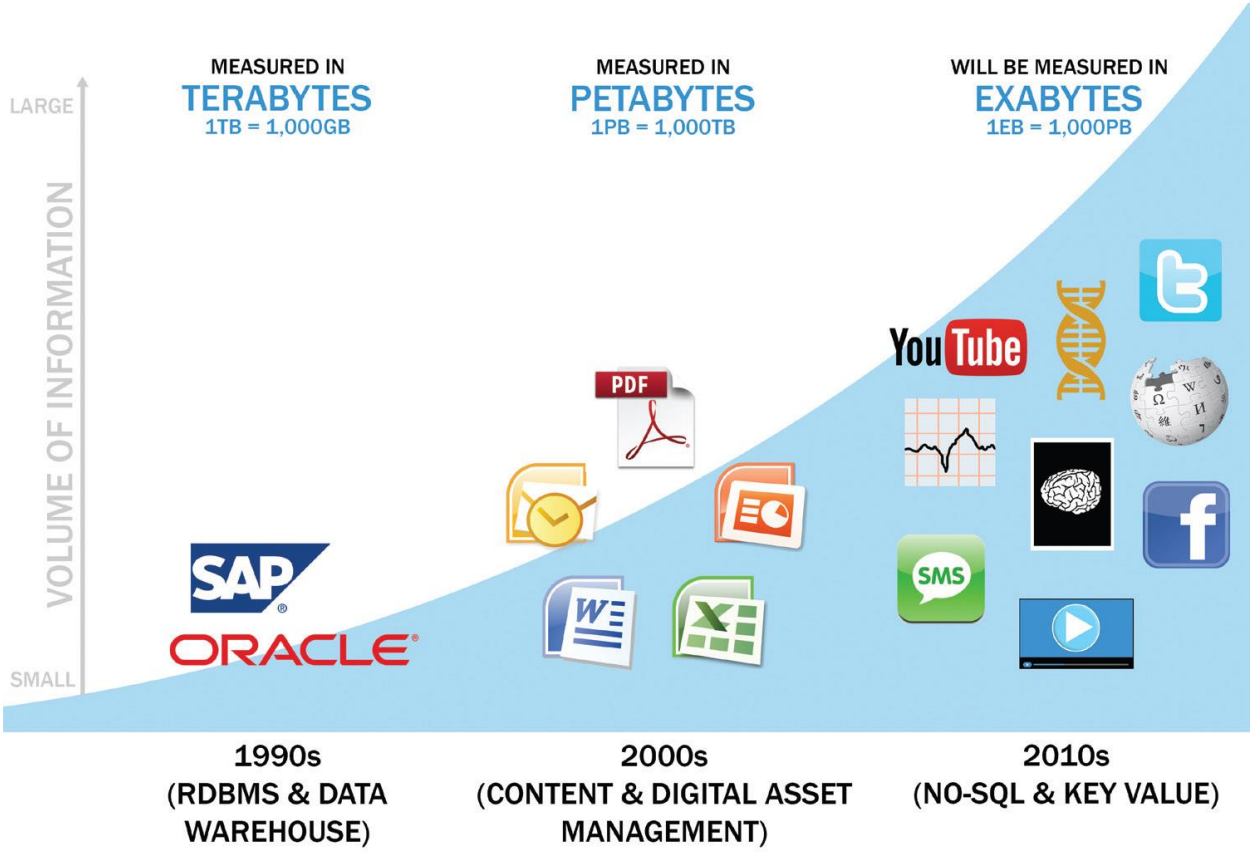
4. At the end of this workflow, analysts get data provisioned for their downstream analytics. Because users generally are not allowed to run custom or intensive analytics on production databases, analysts create data extracts from the EDW to analyze data offline in R or other local analytical tools. Many times these tools are limited to in-memory analytics on desktops analyzing samples of data, rather than the entire population of a dataset. Because these analyses are based on data extracts, they reside in a separate location, and the results of the analysis—and any insights on the quality of the data or anomalies—rarely are fed back into the main data repository.

7 What are the various Drivers of Big Data.

To better understand the market drivers related to Big Data, it is helpful to first understand some past history of data stores and the kinds of repositories and tools to manage these data stores.

As shown in Figure in the 1990s the volume of information was often measured in terabytes. Most organizations analyzed structured data in rows and columns and used relational databases and data warehouses to manage large stores of enterprise information. The following decade saw a proliferation of different kinds of data sources—mainly productivity and publishing tools such as content management repositories and networked attached storage systems—to manage this kind of information, and the data began to increase in size and started to be measured at petabyte scales. In the 2010s, the information that organizations try to manage has broadened to include many other kinds of data. In this era, everyone and everything is leaving a digital footprint. Figure shows a summary perspective on sources of Big Data generated by new applications and the scale and growth rate of the data. These applications, which generate data volumes that can be measured in exabyte scale, provide opportunities for new analytics and driving new value for organizations. The data now comes from multiple sources, such as these:

1.Medical information, such as genomic sequencing and diagnostic imaging
Photos and video footage uploaded to the World Wide Web
2.Video surveillance, such as the thousands of video cameras spread across a city
3.Mobile devices, which provide geospatial location data of the users, as well as metadata about text messages, phone calls, and application usage on smart phones
4.Smart devices, which provide sensor-based collection of information from smart electric grids, smart buildings, and many other public and industry infrastructures
5.Nontraditional IT devices, including the use of radio-frequency identification (RFID) readers, GPS navigation systems, and seismic processing
Data evolution and the rise of Big Data sources

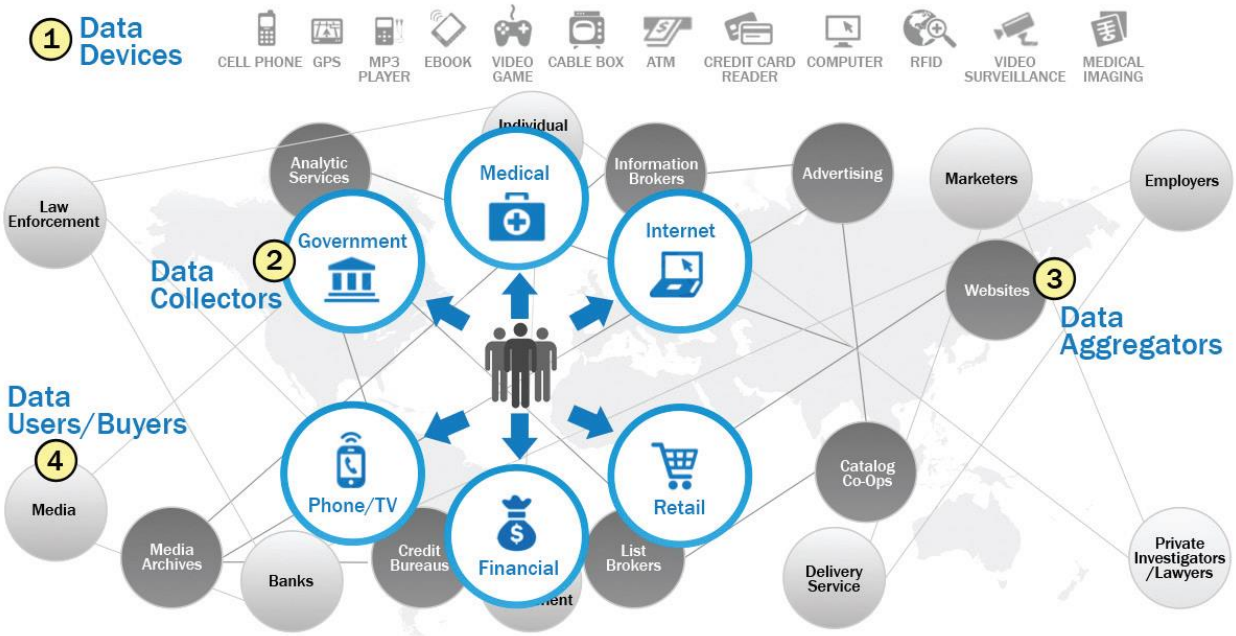


8.Emerging Big Data Ecosystem and a New Approach to Analytics.

Organizations and data collectors are realizing that the data they can gather from individuals contains intrinsic value and, as a result, a new economy is emerging. As this new digital economy continues to evolve, the market sees the introduction of data vendors and data cleaners that use crowdsourcing (such as Mechanical Turk and GalaxyZoo) to test the outcomes of machine learning techniques. Other vendors offer added value by repackaging open source tools in a simpler way and bringing the tools to market. Vendors such as Cloudera, Hortonworks, and Pivotal have provided this value-add for the open source framework Hadoop.

As the new ecosystem takes shape, there are four main groups of players within this interconnected web. These are shown in Figure.

Emerging Big Data ecosystem



1 Data devices [shown in the Figure marked as (1)] and the "Sensornet" gather data from multiple locations and continuously generate new data about this data. For each gigabyte of new data created, an additional petabyte of data is created about that data.

➢ For example, consider someone playing an online video game through a PC, game console, or smartphone. In this case, the video game provider captures data about the skill and levels attained by the player. Intelligent systems monitor and log how and when the user plays the game. As a consequence, the game provider can fine-tune the difficulty of the game, suggest other related games that would most likely interest the user, and offer additional equipment and enhancements for the character based on the user's age, gender, and interests. This information may get stored locally or uploaded to the game provider's cloud to analyze the gaming habits and opportunities for upsell and cross-sell, and identify archetypical profiles of specific kinds of users.

➢ Smartphones provide another rich source of data. In addition to messaging and basic phone usage, they store and transmit data about Internet usage, SMS usage, and real-time location. This metadata can be used for analyzing traffic patterns by scanning the density of smartphones in locations to track the speed of cars or the relative traffic congestion on busy roads. In this way, GPS devices in cars can give drivers real-time updates and offer alternative routes to avoid traffic delays.

➢ Retail shopping loyalty cards record not just the amount an individual spends, but the locations of stores that person visits, the kinds of products purchased, the stores where goods are purchased most often, and the combinations of products purchased together. Collecting this

data provides insights into shopping and travel habits and the likelihood of successful advertisement targeting for certain types of retail promotions.

2 **Data collectors** [the blue ovals, identified within Figure marked as (2)] include sample entities that collect data from the device and users.

- Data results from a cable TV provider tracking the shows a person watches, which TV channels someone will and will not pay for to watch on demand, and the prices someone is willing to pay for premium TV content

- Retail stores tracking the path a customer takes through their store while pushing a shopping cart with an RFID chip so they can gauge which products get the most foot traffic using geospatial data collected from the RFID chips

3 Data aggregators (the dark gray ovals in Figure marked as (3)) make sense of the data collected from the various entities from the "SensorNet" or the "Internet of Things." These organizations compile data from the devices and usage patterns collected by government agencies, retail stores, and websites. In turn, they can choose to transform and package the data as products to sell to list brokers, who may want to generate marketing lists of people who may be good targets for specific ad campaigns.

4 Data users and buyers are denoted by (4) in Figure These groups directly benefit from the data collected and aggregated by others within the data value chain.

- Retail banks, acting as a data buyer, may want to know which customers have the highest likelihood to apply for a second mortgage or a home equity line of credit. To provide input for this analysis, retail banks may purchase data from a data aggregator. This kind of data may include demographic information about people living in specific locations; people who appear to have a specific level of debt, yet still have solid credit scores (or other characteristics such as paying bills on time and having savings accounts) that can be used to infer credit worthiness; and those who are searching the web for information about paying off debts or doing home remodelling projects. Obtaining data from these various sources and aggregators will enable a more targeted marketing campaign, which would have been more challenging before Big Data due to the lack of information or high-performing technologies.

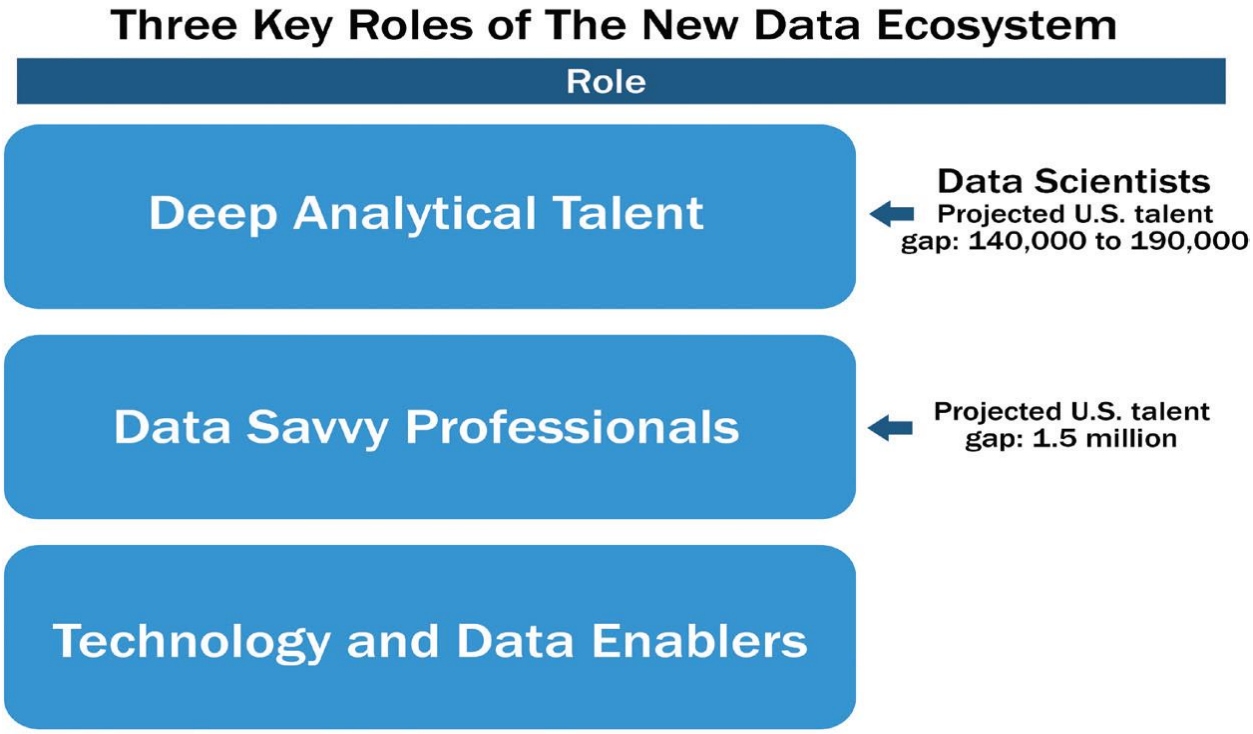
- Using technologies such as Hadoop to perform natural language processing on unstructured, textual data from social media websites, users can gauge the reaction to events such as presidential campaigns. People may, for example, want to determine public sentiments toward a candidate by analysing related blogs and online comments. Similarly, data users may want to track and prepare for natural disasters by identifying which areas a hurricane affects first and how it moves, based on which geographic areas are tweeting about it or discussing it via social media.

9. What are the different Key roles of the new Big Data ecosystem?

The Data Scientist:

New players have emerged to curate, store, produce, clean, and transact data. In addition, the need for applying more advanced analytical techniques to increasingly complex business problems has driven the emergence of new roles, new technology platforms, and new analytical methods.

.



Note: Figures above reflect a projected talent gap in US in 2018, as shown in McKinsey May 2011 article "Big Data: The Next Frontier for Innovation, Competition, and Productivity"

The first group—Deep Analytical Talent— is technically savvy, with strong analytical skills. Members possess a combination of skills to handle raw, unstructured data and to apply complex analytical techniques at massive scales. This group has advanced training in quantitative disciplines, such as mathematics, statistics, and machine learning. To do their jobs, members need access to a robust analytic sandbox or workspace where they can perform large-scale analytical data experiments. Examples of current professions fitting into this group include statisticians, economists, mathematicians, and the new role of the Data Scientist.

The McKinsey study forecasts that by the year 2018, the United States will have a talent gap of 140,000– 190,000 people with deep analytical talent. This does not represent the number of people needed with deep analytical talent; rather, this range represents the difference between what will be available in the workforce compared with what will be needed. In addition, these estimates only reflect forecasted talent shortages in the United States; the number would be much larger on a global basis.

The second group—Data Savvy Professionals—has less technical depth but has a basic knowledge of statistics or machine learning and can define key questions that can be answered

using advanced analytics. These people tend to have a base knowledge of working with data, or an appreciation for some of the work being performed by data scientists and others with deep analytical talent. Examples of data savvy professionals include financial analysts, market research analysts, life scientists, operations managers, and business and functional managers.

The McKinsey study forecasts the projected U.S. talent gap for this group to be 1.5 million people by the year 2018. At a high level, this means for every Data Scientist profile needed, the gap will be ten times as large for Data Savvy Professionals. Moving toward becoming a data savvy professional is a critical step in broadening the perspective of managers, directors, and leaders, as this provides an idea of the kinds of questions that can be solved with data.

The third category of people mentioned in the study is Technology and Data Enablers. This group represents people providing technical expertise to support analytical projects, such as provisioning and administrating analytical sandboxes, and managing large-scale data architectures that enable widespread analytics within companies and other organizations. This role requires skills related to computer engineering, programming, and database administration.

These three groups must work together closely to solve complex Big Data challenges. Most organizations are familiar with people in the latter two groups mentioned, but the first group, Deep Analytical Talent, tends to be the newest role for most and the least understood. For simplicity, this discussion focuses on the emerging role of the Data Scientist. It describes the kinds of activities that role performs and provides a more detailed view of the skills needed to fulfill that role.

There are three recurring sets of activities that data scientists perform:

- Reframe business challenges as analytics challenges. Specifically, this is a skill to diagnose business problems, consider the core of a given problem, and determine which kinds of candidate analytical methods can be applied to solve it.

- Design, implement, and deploy statistical models and data mining techniques on Big Data. This set of activities is mainly what people think about when they consider the role of the Data Scientist: namely, applying complex or advanced analytical methods to a variety of business problems using data.

- Develop insights that lead to actionable recommendations. It is critical to note that applying advanced methods to data problems does not necessarily drive new business value. Instead, it is important to learn how to draw insights out of the data and communicate them effectively.

Data scientists are generally thought of as having five main sets of skills and behavioural characteristics, as shown in Figure

- Quantitative skill: such as mathematics or statistics
- Technical aptitude: namely, software engineering, machine learning, and programming skills
- Sceptical mind-set and critical thinking: It is important that data scientists can examine their work critically rather than in a one-sided way.
- Curious and creative: Data scientists are passionate about data and finding creative ways to solve problems and portray information.
- Communicative and collaborative: Data scientists must be able to articulate the business value in a clear way and collaboratively work with other groups, including project sponsors and key stakeholders. Data scientists are generally comfortable using this blend of skills to acquire, manage, analyse, and visualize data and tell compelling stories about it.

10 What are the different Key Roles for a Successful Analytics Project.

In recent years, substantial attention has been placed on the emerging role of the data scientist. In October 2012, Harvard Business Review featured an article titled "Data Scientist: The Sexiest Job of the 21st Century" , in which experts DJ Patil and Tom Davenport described the new role and how to find and hire data scientists. More and more conferences are held annually focusing on innovation in the areas of Data Science and topics dealing with Big Data. Despite this strong focus on the emerging role of the data scientist specifically, there are actually seven key roles that need to be fulfilled for a high-functioning data science team to execute analytic projects successfully. Figure depicts the various roles and key stakeholders of an analytics project. Each plays a critical part in a successful analytics project. Although seven roles are listed, fewer or more people can accomplish the work depending on the scope of the project, the organizational structure, and the skills of the participants. For example, on a small, versatile team, these seven roles may be fulfilled by only 3 people, but a very large project may require 20 or more people. The seven roles follow.

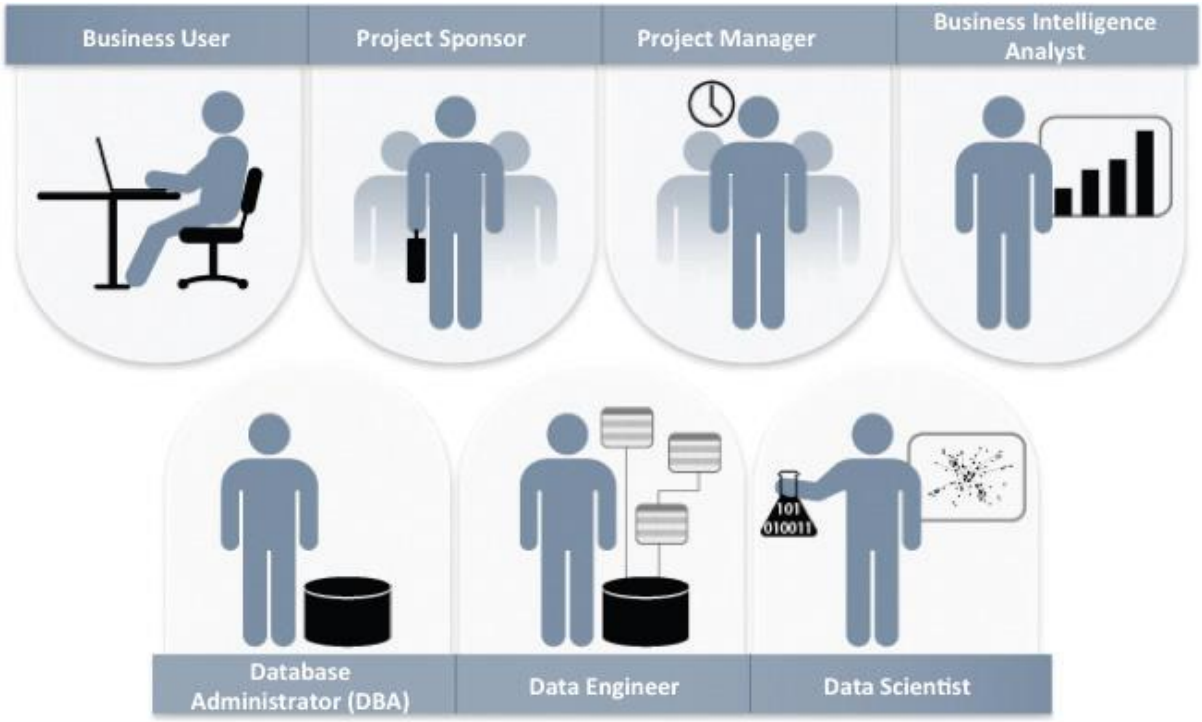


FIGURE *Key roles for a successful analytics project*

Business User: Someone who understands the domain area and usually benefits from the results. This person can consult and advise the project team on the context of the project, the value of the results, and how the outputs will be operationalized. Usually a business analyst, line manager, or deep subject matter expert in the project domain fulfills this role.

Project Sponsor: Responsible for the genesis of the project. Provides the impetus and requirements for the project and defines the core business problem. Generally provides the funding and gauges the degree of value from the final outputs of the working team. This person sets the priorities for the project and clarifies the desired outputs.

Project Manager: Ensures that key milestones and objectives are met on time and at the expected quality.

Business Intelligence Analyst: Provides business domain expertise based on a deep understanding of the data, key performance indicators (KPIs), key metrics, and business intelligence from a reporting perspective. Business Intelligence Analysts generally create dashboards and reports and have knowledge of the data feeds and sources.

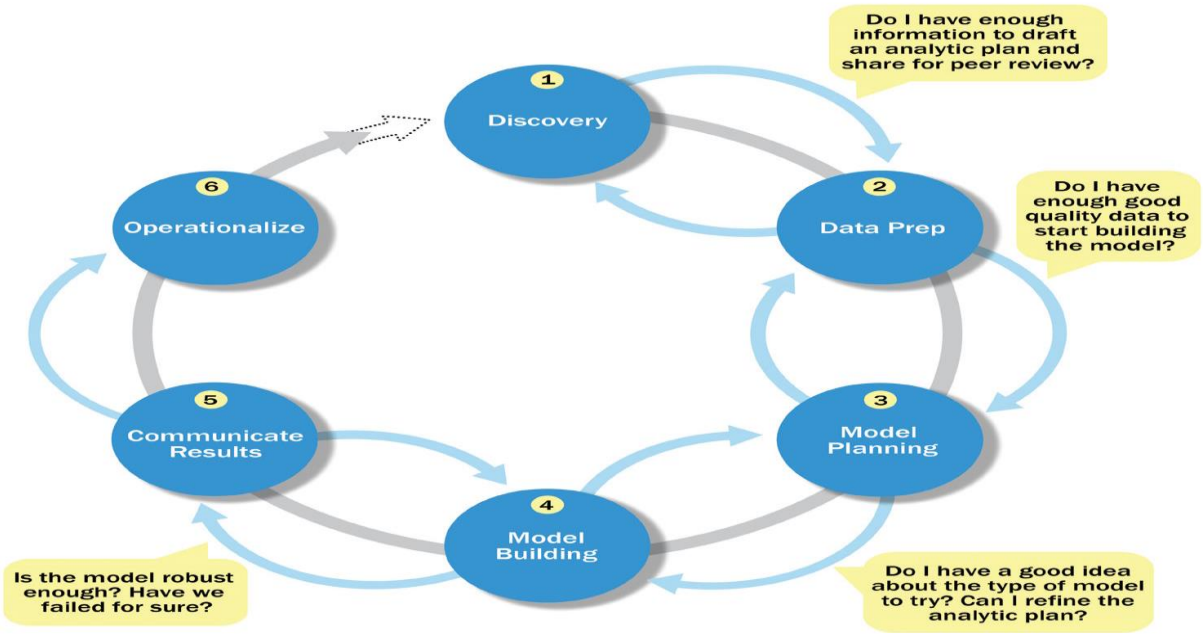
Database Administrator (DBA): Provisions and configures the database environment to support the analytics needs of the working team. These responsibilities may include providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data repositories.

Data Engineer: Leverages deep technical skills to assist with tuning SQL queries for data management and data extraction, and provides support for data ingestion into the analytic sandbox, Whereas the DBA sets up and configures the databases to be used, the data engineer executes the actual data extractions and performs substantial data manipulation to facilitate the analytics. The data engineer works closely with the data scientist to help shape data in the right ways for analyses.

Data Scientist: Provides subject matter expertise for analytical techniques, data modelling, and applying valid analytical techniques to given business problems. Ensures overall analytics objectives are met. Designs and executes analytical methods and approaches with the data available to the project. Although most of these roles are not new, the last two roles—data engineer and data scientist—have become popular and in high demand as interest in Big Data has grown.

11.Explain Background and Overview of Data Analytics Lifecycle:

The Data Analytics Lifecycle is designed specifically for Big Data problems and data science projects. The lifecycle has six phases, and project work can occur in several phases at once. For most phases in the lifecycle, the movement can be either forward or backward. This iterative depiction of the lifecycle is intended to more closely portray a real project, in which aspects of the project move forward and may return to earlier stages as new information is uncovered and team members learn more about various stages of the project. This enables participants to move iteratively through the process and drive toward operationalizing the project work.



Here is a brief overview of the main phases of the Data Analytics Lifecycle:

Phase 1—Discovery: In Phase 1, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn. The team assesses the resources available to support the project in terms of people, technology, time, and data. Important activities in this phase include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data.

Phase 2—Data preparation: Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project. The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox. The ELT and ETL are sometimes abbreviated as ETLT. Data should be

transformed in the ETLT process so the team can work with it and analyse it. In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data

Phase 3—Model planning: Phase 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase. The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.

Phase 4—Model building: In Phase 4, the team develops datasets for testing, training, and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase. The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and workflows (for example, fast hardware and parallel processing, if applicable).

Phase 5—Communicate results: In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

Phase 6—Operationalize: In Phase 6, the team delivers final reports, briefings, code, and technical documents. In addition, the team may run a pilot project to implement the models in a production environment.

12 Describe the Big Data Analytics in Industry Verticals.

Examples of Big Data Analytics in different areas: retail, IT infrastructure, and social media.

Big Data presents many opportunities to improve sales and marketing analytics.

An example of this is the U.S. retailer Target. Charles Duhigg's book *The Power of Habit* discusses how Target used Big Data and advanced analytical methods to drive new revenue. After analyzing consumer purchasing behavior, Target's statisticians determined that the retailer made a great deal of money from three main life-event situations.

- Marriage, when people tend to buy many new products
- Divorce, when people buy new products and change their spending habits
- Pregnancy, when people have many new things to buy and have an urgency to buy them

Target determined that the most lucrative of these life-events is the third situation: pregnancy. Using data collected from shoppers, Target was able to identify this fact and predict which of its shoppers were pregnant. In one case, Target knew a female shopper was pregnant even before her family knew. This kind of knowledge allowed Target to offer specific coupons and incentives to their pregnant shoppers. In fact, Target could not only determine if a shopper was pregnant, but in which month of pregnancy a shopper may be. This enabled Target to manage its inventory, knowing that there would be demand for specific products and it would likely vary by month over the coming nine- to ten-month cycles.

Twitter and Facebook generate massive amounts of unstructured data and use Hadoop and its ecosystem of tools to manage this high volume. Finally, social media represents a tremendous opportunity to leverage social and professional interactions to derive new insights. LinkedIn exemplifies a company in which data itself is the product. Early on, LinkedIn founder Reid

Hoffman saw the opportunity to create a social network for working professionals. As of 2014, LinkedIn has more than 250 million user accounts and has added many additional features and data-related products, such as recruiting, job seeker tools, advertising, and In Maps, which show a social graph of a user's professional network.

13. Discuss Big data analytics Applications.

Big Data is also **data** but with a **huge size**. Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time. In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

Applications of Big Data:

- **Big Data for financial services:** Credit card companies, retail banks, private wealth management advisories, insurance firms, venture funds, and institutional investment banks use big data for their financial services. The common problem among them all is the massive amounts of multi-structured data living in multiple disparate systems which can be solved by big data. Thus big data is used in a number of ways like:
  - Customer analytics
  - Compliance analytics
  - Fraud analytics
  - Operational analytics
  - **Big Data in communications:** Gaining new subscribers, retaining customers, and expanding within current subscriber bases are top priorities for telecommunication service providers. The solutions to these challenges lie in the ability to combine and analyze the masses of customer-generated data and machine-generated data that is being created every day.
- **Big Data for Retail:** Brick and Mortar or an online e-tailer, the answer to staying the game and being competitive is understanding the customer better to serve them. This requires the ability to analyze all the disparate data sources that companies deal with every day, including the weblogs, customer transaction data, social media, store-branded credit card data, and loyalty program data.

Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers. In his report *Big Data in Big Companies*, IIA Director of Research Tom Davenport interviewed more than 50 businesses to understand how they used big data. He found they got value in the following ways:

1. **Cost reduction.** Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.
2. **Faster, better decision making.** With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they've learned.

3. **New products and services.** With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs.

4. **Life Sciences.** Clinical research is a slow and expensive process, with trials failing for a variety of reasons. Advanced analytics, artificial intelligence (AI) and the Internet of Medical Things (IoMT) unlocks the potential of improving speed and efficiency at every stage of clinical research by delivering more intelligent, automated solutions

5. **Banking.**Financial institutions gather and access analytical insight from large volumes of unstructured data in order to make sound financial decisions. Big data analytics allows them to access the information they need when they need it, by eliminating overlapping, redundant tools and systems.

6. Manufacturing. For manufacturers, solving problems is nothing new. They wrestle with difficult problems on a daily basis - from complex supply chains, to motion applications, to labor constraints and equipment breakdowns. That's why big data analytics is essential in the manufacturing industry, as it has allowed competitive organizations to discover new cost saving opportunities and revenue opportunities.

7. **Health Care.**Big data is a given in the health care industry. Patient records, health plans, insurance information and other types of information can be difficult to manage – but are full of key insights once analytics are applied. That's why big data analytics technology is so important to heath care. By analyzing large amounts of information – both structured and unstructured – quickly, health care providers can provide lifesaving diagnoses or treatment options almost immediately.

8. **Government.**Certain government agencies face a big challenge: tighten the budget without compromising quality or productivity. This is particularly troublesome with law enforcement agencies, which are struggling to keep crime rates down with relatively scarce resources. And that's why many agencies use big data analytics; the technology streamlines operations while giving the agency a more holistic view of criminal activity. Retail.Customer service has evolved in the past several years, as savvier shoppers expect retailers to understand exactly what they need, when they need it. Big data analytics technology helps retailers meet those demands. Armed with endless amounts of data from customer loyalty programs, buying habits and other sources, retailers not only have an in-depth understanding of their customers, they can also predict trends, recommend new products – and boost profitability.

14. Define the term Big Data Analytics and explain with a use case.

Big Data Analytics: Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. With today's technology, it's possible to analyse your data and get answers from it almost immediately – an effort that's slower and less efficient with more traditional business intelligence solutions.

Manufacturing Big Data Use Cases

The digital revolution has transformed the manufacturing industry. Manufacturers are now finding new ways to harness all the data they generate to improve operational efficiency, streamline business processes, and uncover valuable insights that will drive profits and growth.

Predictive Maintenance: Big data can help predict equipment failure. Potential issues can be discovered by analyzing both structured data (equipment year, make, and model) and multistructured data (log entries, sensor data, error messages, engine temperature, and other factors). With this data, manufacturers can maximize parts and equipment uptime and deploy maintenance more cost effectively.

This data can be used to predict more than just equipment failure. For many manufacturing processes, it's also important to predict the remaining optimal life of systems and components to ensure that they perform within specifications. Falling out of tolerance—even if nothing is broken—can be as bad as failure. For example: in drug manufacturing a faulty, but still functional, component could introduce too much or too little of the active ingredient.

Challenges: Companies must integrate data coming from different formats and identify the signals that will lead to optimizing maintenance.

OperationalEfficiency: Operational efficiency is one of the areas in which big data can have the most impact on profitability. With big data, you can analyse and assess production processes, proactively respond to customer feedback, and anticipate future demands.

Challenges: Data teams must balance the data volume with the growing number of sources, users, and applications.

ProductionOptimization: Optimizing production lines can decrease costs and increase revenue. Big data can help manufacturers understand the flow of items through their production lines and see which areas can benefit. Data analysis will reveal which steps lead to increased production time and which areas are causing delays.

Challenges
Optimizing production requires manufacturers to analyze their production equipment data, material use, and other factors. Combining the different kinds of data can pose a challenge.
Retail Big Data Use Cases

Competition is fierce in retail. To stay ahead, companies strive to differentiate themselves. Big data is being used across all stages of the retail process—from product predictions to demand forecasting to in-store optimization. Using big data, retailers are finding new ways to innovate.

Product Development: Big data can help you anticipate customer demand. By classifying key attributes of past and current products and then modelling the relationship between those attributes and the commercial success of the offerings, you can build predictive models for new products and services. Dig deeper by using the data and analytics from focus groups, social media, test markets, and early store rollouts to plan, produce, and launch new products.

Challenges: Companies will have to analyse what can be a high volume of data coming in varying formats, and then create segments according to customer behaviour. They will also have to identify sophisticated use patterns and behaviour and map them to potential new offerings.

Customer Experience: The race for customers is on. Big data provides retailers with a clearer view of the customer experience that they can use to fine-tune their operations. By gathering data from social media, web visits, call logs and other company interactions, and other data sources, companies can improve customer interactions and maximize the value delivered. Big data analytics can be used to deliver personalized offers, reduce customer churn, and proactively handle issues.

Challenges: Integrating a high volume of data from various sources can be difficult. Once the data is integrated, path analysis can be used to identify experience paths and correlate them with various sets of behaviour.

Customer Lifetime Value: All customers are valuable. But some are more valuable than others. Big data provides you with insights on customer behaviour and spending patterns so you can identify your best customers. Once you know who they are, marketing can target them

with special offers. Sales teams can devote more time to them. Customer service can work more proactively if it appears they may leave.

Challenges: To identify your high-value customers, you will need to analyse a high volume of customer transaction data and create sophisticated models that examine past behaviour and predict future actions.

The In-Store Shopping Experience: Big data can be used to improve the in-store experience. Many retailers are starting to analyse data from mobile apps, in-store purchases, and geolocations to optimize merchandizing encourage customers to complete purchases.

Challenges: Complex graphs and path analyses are required to identify customer paths and behaviour. This data must then be correlated and joined with multiple datasets to correctly analyse store behaviour.

Pricing Analytics and Optimization: Retailers need to know the true profitability of their customers, how markets can be segmented, and the potential of any future opportunities. End-to-end profit and margin analysis can help with identifying pricing improvement opportunities and areas where profits may be leaking.

Challenges: To correctly analyse pricing data, retailers need to manage millions of pieces of transaction data and work with many different kinds of data sets.

Healthcare Big Data Use Cases

Healthcare organizations are using big data for everything from improving profitability to helping save lives. Healthcare companies, hospitals, and researchers collect massive amounts of data. But all of this data isn't useful in isolation. It becomes important when the data is analysed to highlight trends and threats in patterns and create predictive models.

Genomic Research: Big data can play in a significant role in genomic research. Using big data, researchers can identify disease genes and biomarkers to help patients pinpoint health issues they may face in the future. The results can even allow healthcare organizations to design personalized treatments.

Challenges: Companies must integrate data coming from different formats and identify the signals that will lead to optimizing maintenance.

Patient Experience and Outcomes: Healthcare organizations seek to provide better treatment and improved quality of care—without increasing costs. Big data helps them improve the patient experience in the most cost-efficient manner. With big data, healthcare organizations can create a 360-degree view of patient care as the patient moves through various treatments and departments.

Challenges: Improving the patient experience requires a large volume of patient data, some of which could be multistructured data, such as doctor notes or images. Additionally, to analyse patient journeys, path and graph analyses are often needed.

Claims Fraud: For every healthcare claim, there can be hundreds of associated reports in a variety of different formats. This makes it extremely difficult to verify the accuracy of insurance incentive programs and find the patterns that indicate fraudulent activity. Big data helps healthcare organizations detect potential fraud by flagging certain behaviours for further examination.

Challenges: Claims fraud analytics is a complex process that involves integrating different data sets, analysing the claims data, and identifying complex fraud patterns.

Healthcare Billing Analytics: Big data can improve the bottom line. By analysing billing and claims data, organizations can discover lost revenue opportunities and places where payment cash flows can be improved. This use case requires integrating billing data from various payers, analysing a large volume of that data, and then identifying activity patterns in the billing data.

Challenges: Sifting through large volumes of data can be complicated, especially when it comes to integrating different data sources.

Oil and Gas Big Data Use Cases

For the past few years, the oil and gas industry has been leveraging big data to find new ways to innovate. The industry has long made use of data sensors to track and monitor the performance of oil wells, machinery, and operations. Oil and gas companies have been able to

harness this data to monitor well activity, create models of the Earth to find new oil sources, and perform many other value-added tasks.

Predictive Equipment Maintenance: Oil and gas companies often lack visibility into the condition of their equipment, especially in remote offshore and deep-water locations. Big data can help by providing insight so companies can predict the remaining optimal life of their systems and components, ensuring that their assets operate at optimum production efficiency.

Challenges: Machine, log, and sensor data from different types of equipment comes in varying formats. Integrating all of this data can be difficult. Moreover, the data needs to be analysed quickly and put into operation to effectively prevent downtime.

Oil Exploration and Discovery: Exploring for oil and gas can be expensive. But companies can make use of the vast amount of data generated in the drilling and production process to make informed decisions about new drilling sites. Data generated from seismic monitors can be used to find new oil and gas sources by identifying traces that were previously overlooked.

Challenges: To discover potential new oil deposits, companies will need to integrate and analyse an enormous volume of unstructured data.

Oil Production Optimization: Unstructured sensor and historical data can be used to optimize oil well production. By creating predictive models, companies can measure well production to understand usage rates. With deeper data analysis, engineers can determine why actual well outputs aren't tallying with their predictions.

Challenges: This use case involves analysing a large volume of data. Complex algorithms are also needed to identify the curve shape associated with that data to identify trends.

Telecommunications Big Data Use Cases

The popularity of smart phones and other mobile devices has given telecommunications company's tremendous growth opportunities. But there are challenges as well, as organizations work to keep pace with customer demands for new digital services while managing an ever-expanding volume of data.

Optimize Network Capacity: Optimal network performance is essential for a telecom's success. Network usage analytics can help companies identify areas with excess capacity and reroute bandwidth as needed. Big data analytics can help them plan for infrastructure investments and design new services that meet customer demands. With new insights, telecoms are able maintain customer loyalty and avoid losing revenue to competitors.

Challenges: In addition to creating complex models of relationships between network services and customers, network usage analytics requires analysing a high volume of call detail records.

Telecom Customer Churn: By analysing the data telecoms already have about service quality, convenience, and other factors, telecoms can predict overall customer satisfaction. And they can set up alerts when customers are at risk of churning—and take action with retention campaigns and proactive offers.

Challenges: This use case requires analysing past and current data to create a new model to predict churn, which can be done with time-series and relational analytics to identify patterns and behaviour. Graph analytics helps identify relationships between customers who have recently churned and current customers who may be more likely to churn because they know someone who has churned.

New Product Offerings: Unstructured sensor and historical data can be used to optimize oil well production. By creating predictive models, companies can measure well production to understand usage rates. With deeper data analysis, engineers can determine why actual well outputs aren't tallying with their predictions.

Challenges
This use case involves analysing a large volume of data. Complex algorithms are also needed to identify the curve shape associated with that data to identify trends.

Financial Services Big Data Use Cases

Forward-thinking banks and financial services firms are capitalizing on big data. From capturing new market opportunities to reducing fraud, financial services organizations have been able to convert big data into a competitive advantage.

Fraud and Compliance: When it comes to security, it's not just a few rogue hackers. The financial services industry is up against entire expert teams. While security landscapes and compliance requirements are constantly evolving. Using big data, companies can identify patterns that indicate fraud and aggregate large volumes of information to streamline regulatory reporting.

Challenges: Collecting and aggregating disparate data sources can be difficult.

Anti-Money Laundering: Financial services firms are under more pressure than ever before from governments passing anti-money laundering laws. These laws require that banks show proof of proper diligence and submit suspicious activity reports. In this extraordinarily complicated arena, big data analytics can help companies identify potential fraud patterns.

Challenges: This use case requires analysing large volumes of transaction data (which can include structured and multi-structured data) and then identifying complex AML transactions. In addition, graph analytics will reveal the hidden relationships.

Financial Regulatory and Compliance Analytics: Financial services companies must be in compliance with a wide variety of requirements concerning risk, conduct, and transparency. At the same time, banks must comply with the Dodd-Frank Act, Basel III, and other regulations that require detailed reporting.

Challenges: Financial services companies must bring together a large volume of data, create advanced risk models, and do this quickly without adversely affecting other projects.

15. How the data is prepared in the Data Analytics Life Cycle?

The second phase of the Data Analytics Lifecycle involves data preparation, which includes the steps to explore, pre-process, and condition data prior to modelling and analysis. In this phase,

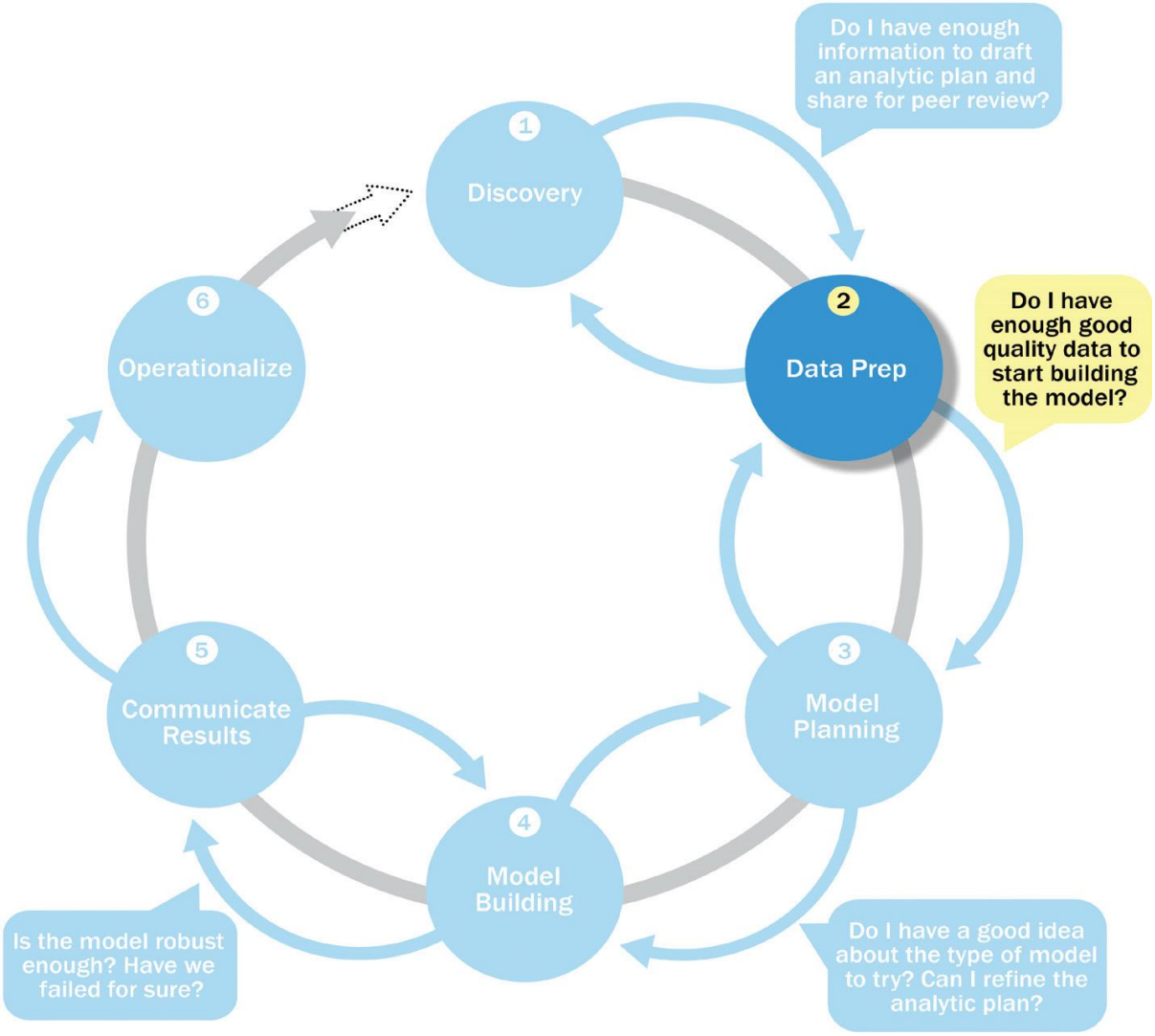
Preparing the Analytic Sandbox: the team needs to create a robust environment in which it can explore the data that is separate from a production environment. Usually, this is done by preparing an analytics sandbox.

Performing ETLT: To get the data into the sandbox, the team needs to perform ETLT, by a combination of extracting, transforming, and loading data into the sandbox. Once the data is in the sandbox, the team needs to learn about the data and become familiar with it.

Learning about the Data: Understanding the data in detail is critical to the success of the project. The team also must decide how to condition and transform data to get it into a format to facilitate subsequent analysis.

Survey and Visualize: The team may perform data visualizations to help team members understand the data, including its trends, outliers, and relationships among data variables. Data preparation tends to be the most labour-intensive step in the analytics lifecycle. In fact, it is common for teams to spend at least 50% of a data science project's time in this critical phase. If the team cannot obtain enough data of sufficient quality, it may be unable to perform the subsequent steps in the lifecycle process. Figure shows an overview of the Data Analytics Lifecycle for Phase 2. The data preparation phase is generally the most iterative and the one that teams tend to underestimate most often. This is because most teams and leaders are anxious to begin analyzing the data, testing hypotheses, and getting answers to some of the questions

posed in Phase 1. Many tend to jump into Phase 3 or Phase 4 to begin rapidly developing models and algorithms without spending the time to prepare the data for modeling. Consequently, teams come to realize the data they are working with does not allow them to execute the models they want, and they end up back in Phase 2 anyway.



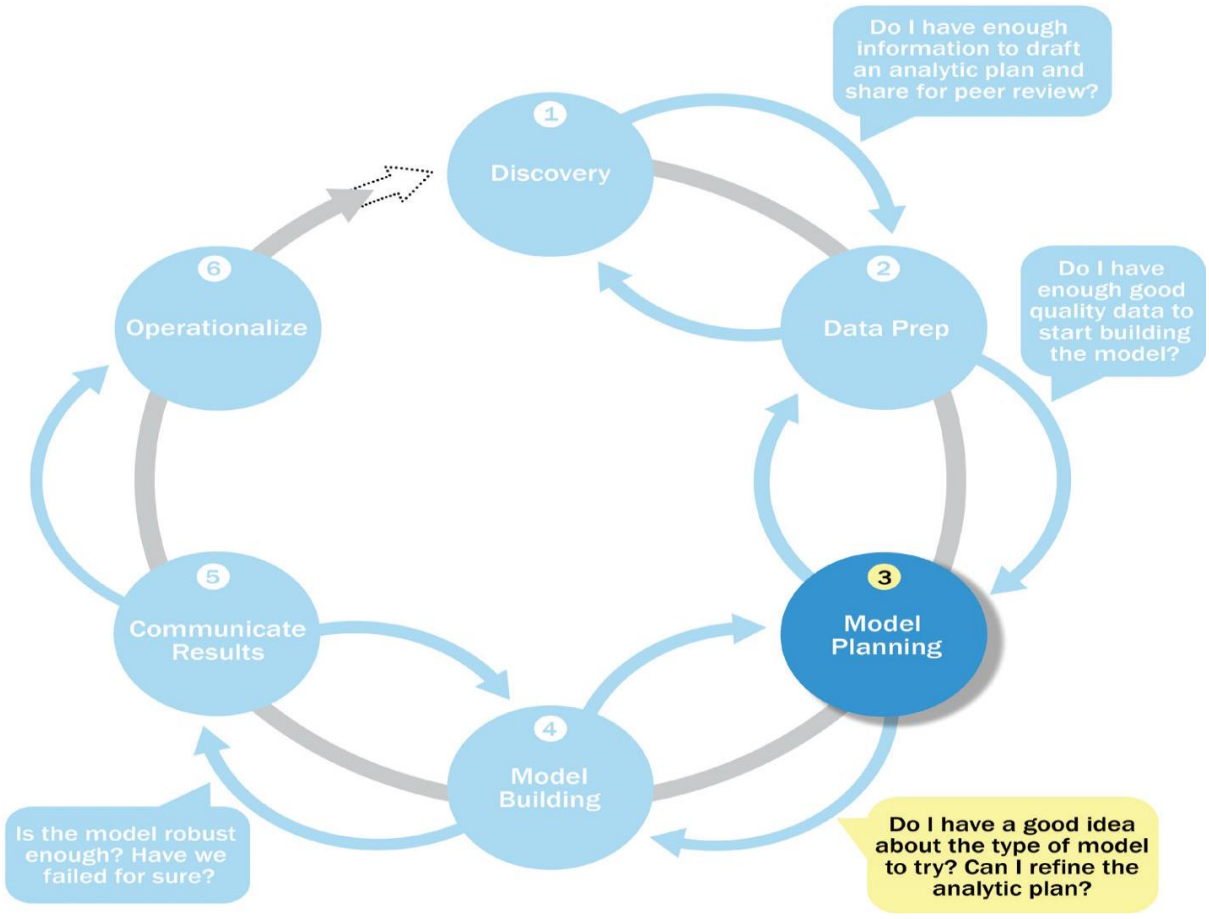
16. **How the model is build and planned in the Data Analytics Life Cycle?**

Model Planning:

In Phase 3, the data science team identifies candidate models to apply to the data for clustering, classifying, or finding relationships in the data depending on the goal of the project, as shown in Figure. It is during this phase that the team refers to the hypotheses developed in Phase 1, when they first became acquainted with the data and understanding the business problems or domain area. These hypotheses help the team frame the analytics to execute in Phase 4 and select the right methods to achieve its objectives. Some of the activities to consider in this phase include the following:

● Assess the structure of the datasets. The structure of the datasets is one factor that dictates the tools and analytical techniques for the next phase. Depending on whether the team plans to analyse textual data or transactional data, for example, different tools and approaches are required.

● Ensure that the analytical techniques enable the team to meet the business objectives and accept or reject the working hypotheses. Determine if the situation warrants a single model or a series of techniques as part of a larger analytic workflow. A few example models include association rules and logistic regression. Other tools, such as Alpine Miner, enable users to set up a series of steps and analyses and can serve as a front-end user interface (UI) for manipulating Big Data sources in PostgreSQL.



Model Building

In Phase 4, the data science team needs to develop datasets for training, testing, and production purposes. These datasets enable the data scientist to develop the analytical model and train it ("training data"), while holding aside some of the data ("hold-out data" or "test data") for testing the model. During this process, it is critical to ensure that the training and test datasets are sufficiently robust for the model and analytical techniques. A simple way to think of these datasets is to view the training dataset for conducting the initial experiments and the test sets for validating an approach once the initial experiments and models have been run. In the model building phase, shown in Figure. An analytical model is developed and fit on the training data

and evaluated (scored) against the test data. The phases of model planning and model building can overlap quite a bit, and in practice one can iterate back and forth between the two phases for a while before settling on a final model. Although the modelling techniques and logic required to develop models can be highly complex,

The Actual duration of this phase can be short compared to the time spent preparing the data and defining the approaches. In general, plan to spend more time preparing and learning the data (Phases 1–2) and crafting a presentation of the findings (Phase 5). Phases 3 and 4 tend to move more quickly, although they are more complex from a conceptual standpoint.

As part of this phase, the data science team needs to execute the models defined in Phase 3.

During this phase, users run models from analytical software packages, such as R or SAS, on file extracts and small datasets for testing purposes. On a small scale, assess the validity of the model and its results. For instance, determine if the model accounts for most of the data and has robust predictive power. At this point, refine the models to optimize the results, such as by modifying variable inputs or reducing correlated variables where appropriate. In Phase 3, the team may have had some knowledge of correlated variables or problematic data attributes, which will be confirmed or denied once the models are actually executed.

When immersed in the details of constructing models and transforming data, many small decisions are often made about the data and the approach for the modelling. These details can be easily forgotten once the project is completed. Therefore, it is vital to record the results and logic of the model during this phase. In addition, one must take care to record any operating assumptions that were made in the modelling process regarding the data or the context.